Autonomous Vocal and Backing Track Mixing

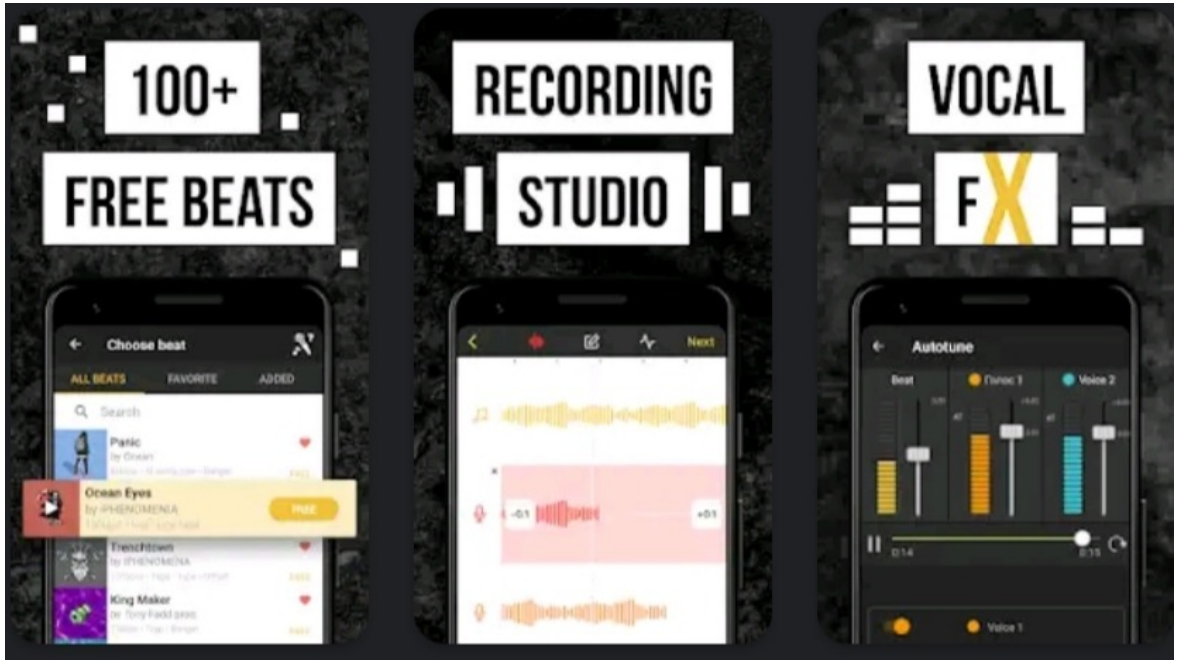
Kelian (Mike) Li

Music Informatics Group

Georgia Tech Center for Music Technology
College of Design

Motivation

- Karaoke apps
- Amateur music makers



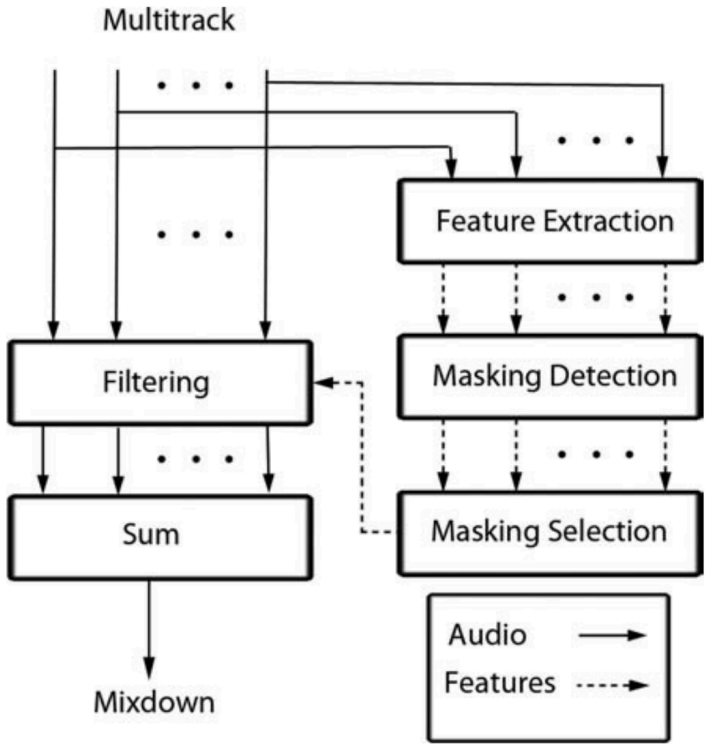
Baseline System: level and compression

Use the average values extracted from the source-separated Million Song Dataset

- Level balance
  - -1.77 dB vocal-to-backing track ratio
- Compression
  - 16.4 dB loudness range

Baseline System: EQ

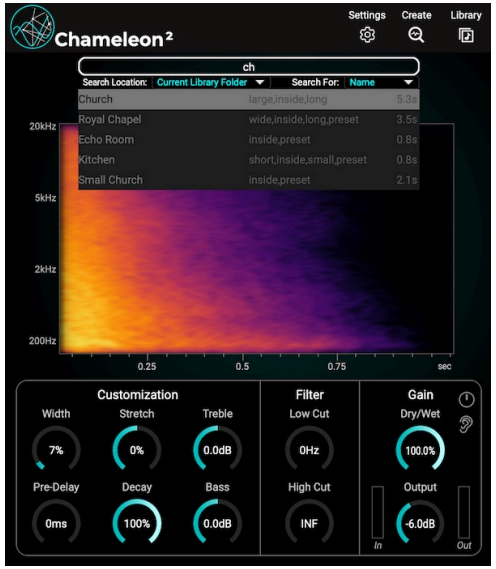
- EQ
  - Frequency unmasking



Rule-based System: reverb

1. Get the estimated impulse responses from the Chameleon plugin
2. Estimate the reverb parameters by the genetic algorithm
3. Use mean values extracted from the MUSDB18 train set

- **Reverb**
  - Dry/wet ratio: 11.5%
  - Reverb time: Linear mapping from tempo
  - Room size: 14.54
  - Fade in time: 0.68 s



Georgia Tech Center for Music Technology
College of Design

Data-driven System

Train a convolutional neural network to predict direct or intermediate mixing parameters based on the input audio

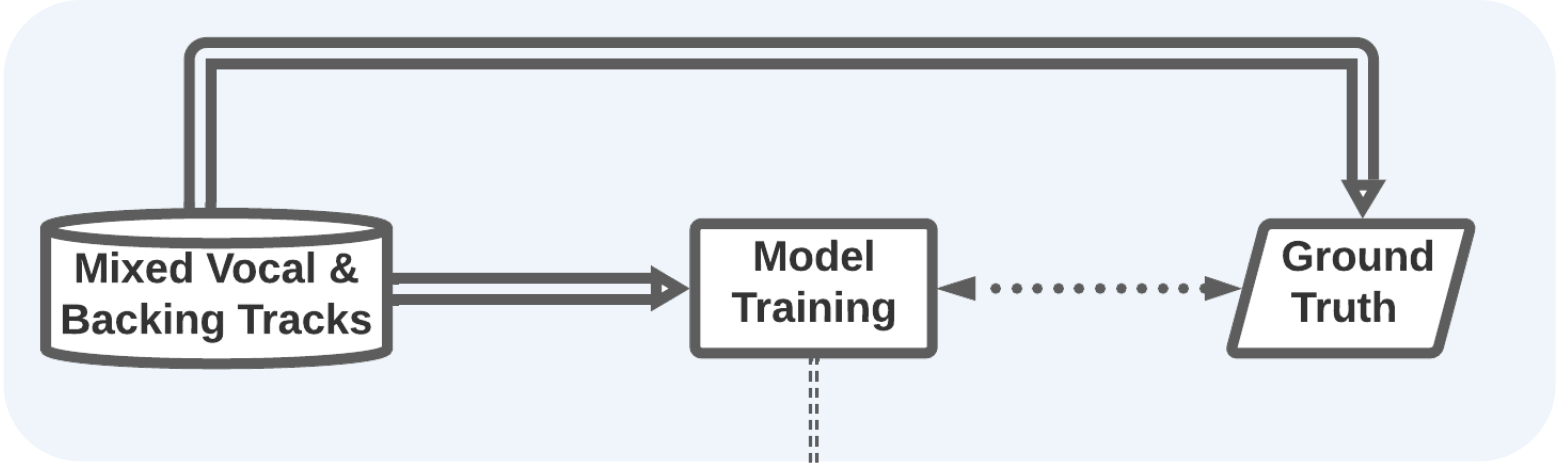
- model input:

  Mel-spectrogram of the vocal and the backing track

Data-driven System

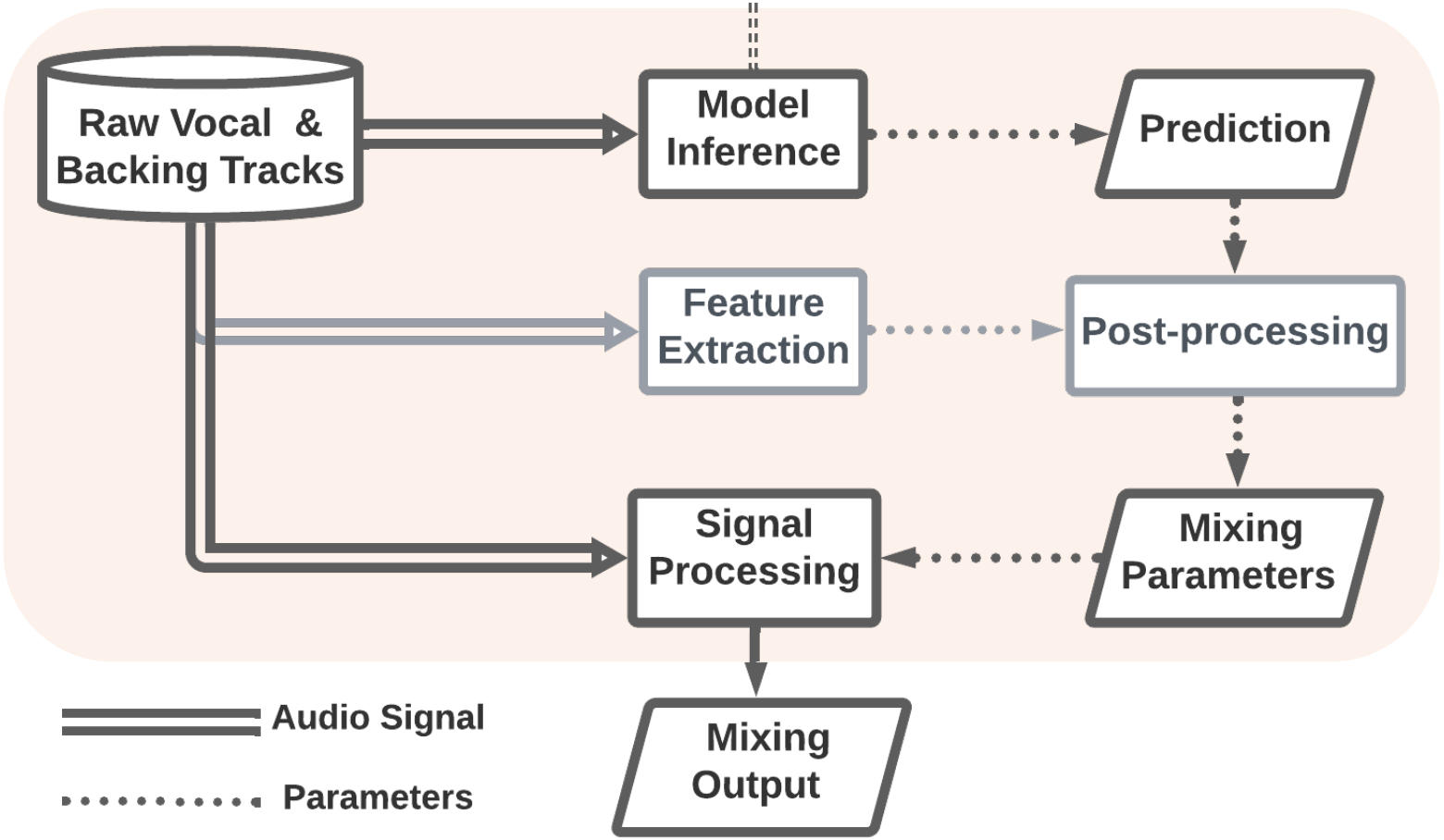
Conv2d (input channels = 2, output channels = 8)
BatchNorm
ReLU
MaxPool2d
Conv2d (input channels = 8, output channels = 16)
BatchNorm
ReLU
MaxPool2d
Conv2d (input channels = 16, output channels = 32)
BatchNorm
ReLU
MaxPool2d
Conv2d(input channels = 32, output channels = 64)
BatchNorm
ReLU
MaxPool2d
Dropout (p = 0.3)
MLP (input features = 768)

Model Training



Ground truth is the direct or intermediate mixing parameters

Model Inference

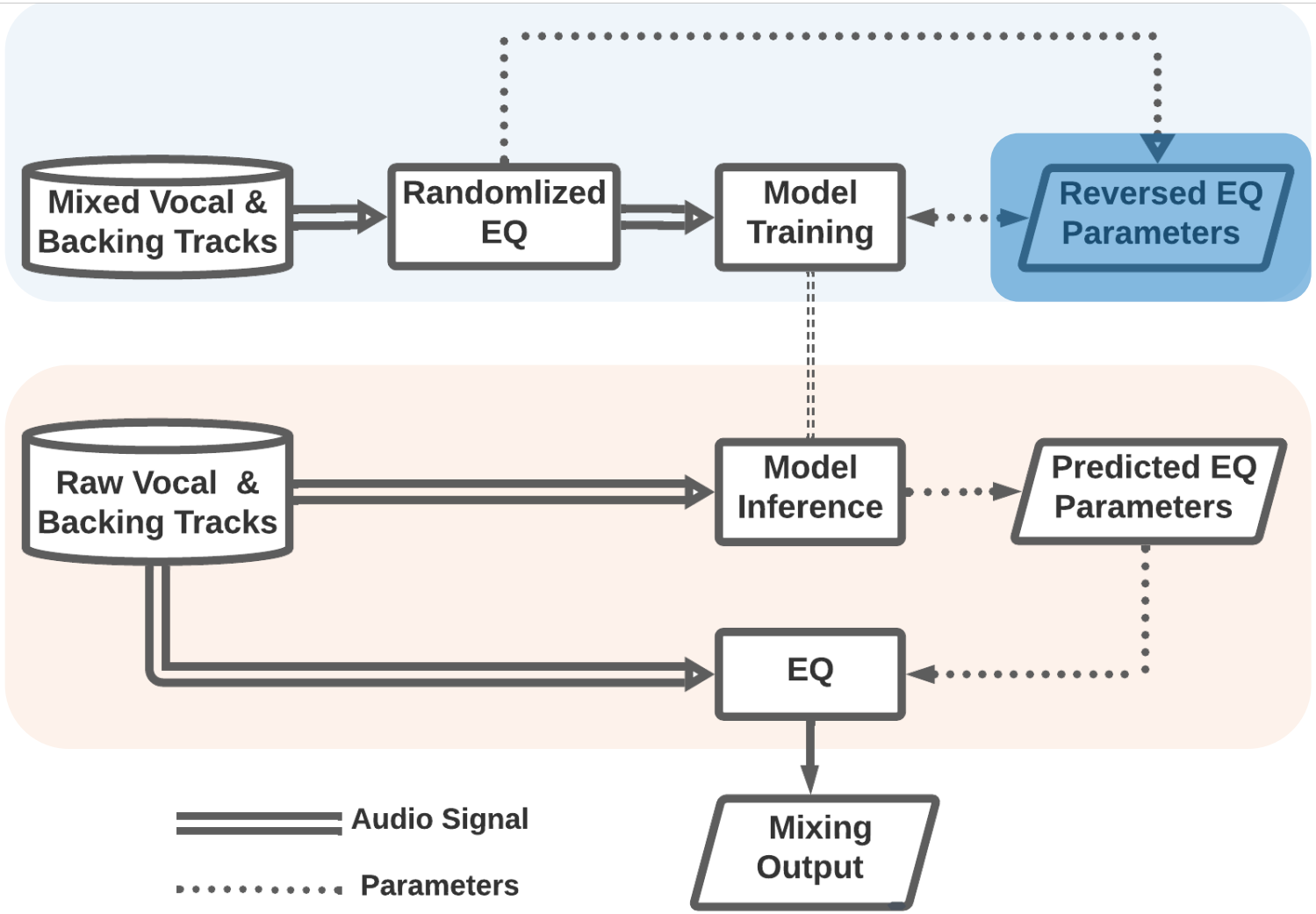


Optionally, convert intermediate model outputs into direct mixing parameters

Data-driven System: EQ

Model Training

Model Inference



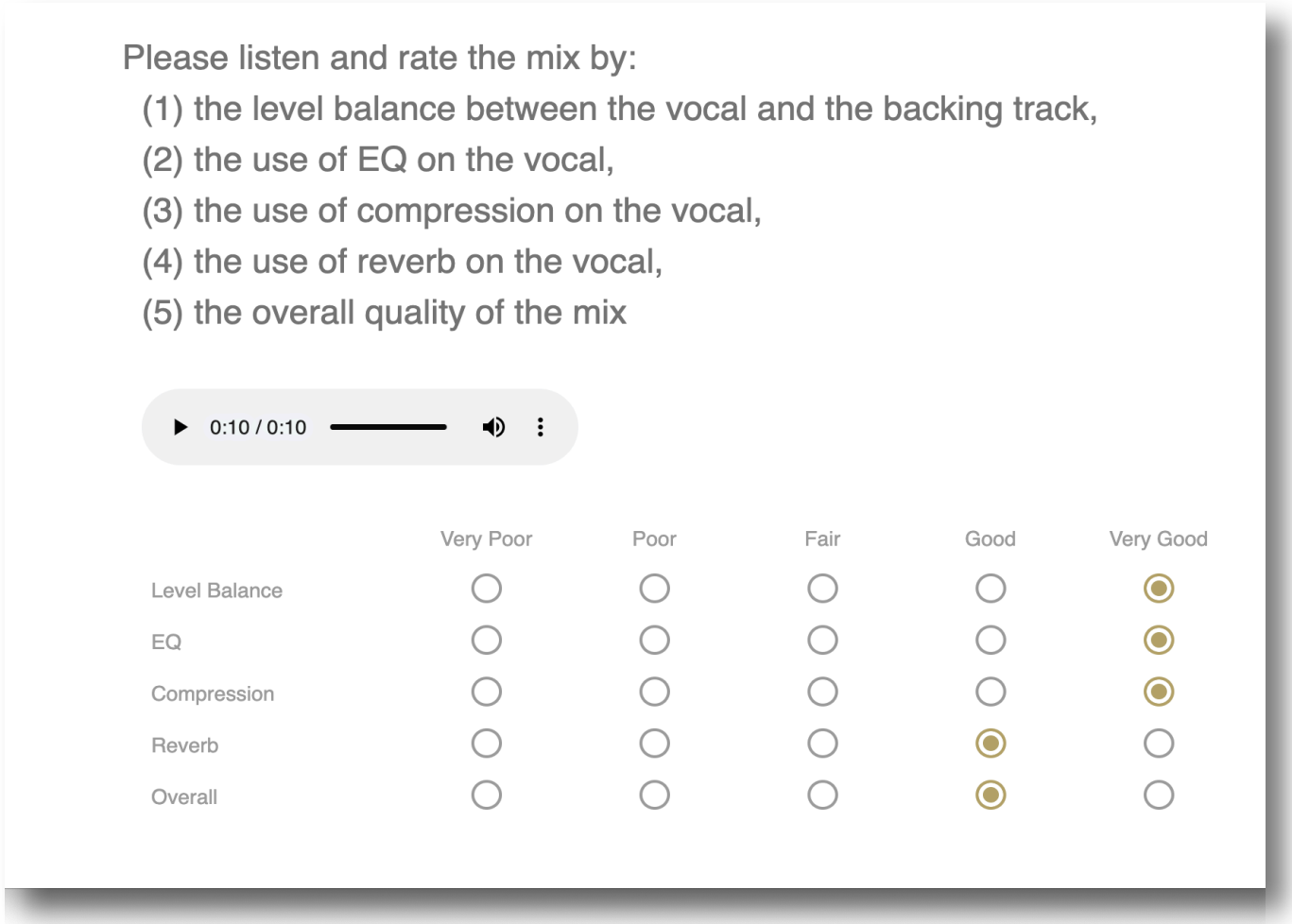
If the mixed vocal is boosted at a center frequency, we should learn to cut at that frequency.

Objective Evaluation

Validation on the MUSDB test set, 48 songs in total

	relative loudness (dB)	loudness range (dB)	EQ gain (dB)	dry/wet ratio (%)	reverb time (s)	room size	fade-in time (s)
CNN	**1.64**	**2.63**	4.48	**6.13**	**1.006**	**7.30**	**0.401**
mean	2.13	2.88	**3.33**	7.16	1.007	7.31	0.403

Listening Test



[Short link to the survey]

[QR code]

Georgia Tech Center for Music Technology
College of Design